# STRUCTURAL GRAMMAR OF THE VOYNICH MANUSCRIPT: CHARACTERIZATION AND MECHANISM DISCRIMINATION

A PREPRINT

**Asa Shepard**
Williams College
as66@williams.edu

April 19, 2026

**ABSTRACT**

This paper has two contributions. First, it documents and evaluates a methodology for autonomous long-running research agents: a multi-skill loop in which a primary agent runs experiments under explicit failure-mode constraints while a higher-capability meta-review agent audits at phase boundaries, with persistent state files acting as the agent's memory and a separate human-edited commitment ledger gating which findings become public claims. Second, it presents the case study that generated this evaluation: a 579-experiment autonomous investigation of the Voynich Manuscript that produced five replicated structural-grammar findings (a three-layer positional architecture, a complete A/B-dialect terminal-layer inversion, a $7\times7$ onset-class bigram grammar with cross-dialect invariants, a directed paragraph-level succession chain, and a systematic dialect-marking suffix transformation), a mechanism-discrimination result narrowing the space of viable generative hypotheses, and a constructive existence proof of the surviving mechanism class. We frame the structural findings as evidence the methodology can produce publishable research, and we propose four falsifiable hypotheses about when long-running research agents work, including that external meta-review catches errors the primary agent's self-checking misses, that commitment ledgers prevent claim drift over long horizons, and that pre-registration with explicit "unfair-test" disclosure reduces overreach errors. Five substantive corrections caught during the run support these hypotheses; we report each in detail.

## 1 Introduction

Autonomous AI agents can now run long-horizon research projects: design experiments, execute them, interpret results, generate follow-up hypotheses, and accumulate state across hundreds of iterations. What is unclear is when the resulting work is trustworthy, what failure modes recur, and what scaffolding makes the difference between productive autonomy and confidently-wrong output.

This paper proposes a methodology and offers one full project run as evidence. The methodology is a multi-skill autonomous research loop with a specific architectural shape: an explicit identity document, five domain-tailored procedural skills, a persistent state-file system that acts as the agent's memory, a separate human-edited commitment ledger gating which internal findings become public claims, mandatory pre-registration with "unfair-test" disclosure, and a higher-capability meta-review agent invoked at phase boundaries to catch errors the primary agent's self-checking misses. We hypothesize that this architecture extends usefully beyond the present case study and propose four specific, falsifiable claims about when it works.

The case study is the Voynich Manuscript, a 15th-century undeciphered codex (Beinecke MS 408, radiocarbon-dated 1404–1438 CE (Hodgins et al., 2011)) containing roughly 38,000 tokens that has resisted analysis for over a century. We chose it not because its contents are practically important but because it has the right shape for evaluating an autonomous research agent: a long-running tradition of computational analysis to anchor against (so we can tell rediscovery from novelty), a clear empirical substrate (the manuscript text), no possibility of post-hoc validation by speaking to an author

or running a controlled physical experiment, and a high prior on positive results being methodological artifacts. An agent that makes real progress in this domain must do so by ruling things out, not by stumbling into the right answer.

Three broad generative hypotheses dominate the prior computational literature on Voynichese: that it encodes natural language through a cipher (with the Naibbe homophonic cipher of Greshko (2025) as the most recent specific proposal); that it is meaningless text generated by a mechanical procedure (with the self-citation algorithm of Timm and Schinner (2020) as the most developed); or that it is a structured system, such as a constructed language or mnemonic code, whose regularities reflect intrinsic rules. The structural findings the agent produced over $579$ experiments are presented in Sections 4–5: a three-layer positional architecture with complete A/B-dialect terminal-layer inversion, a $7\times7$ bigram succession matrix with cross-dialect invariants, a directed paragraph-level succession chain that no null model reproduces at the $100$th percentile, a systematic dialect-marking suffix transformation, and a mechanism-discrimination result establishing a dual constraint that no position-blind or substitution-only mechanism satisfies. All five characterization findings replicate under three independent transliteration systems against pre-registered thresholds. We report these as evidence the methodology can produce a publishable result; whether they advance the Voynich literature is a secondary matter for that community.

We make a methodological disclosure prominently. The work was conducted by an autonomous multi-agent loop. A Sonnet-class primary agent ran the experimental loop continuously; an Opus-class meta-review agent was invoked at phase boundaries to perform independent audits. All strategic decisions, the system's architectural definitions (the identity document, the five skills, the commitment ledger), and the writing of this paper are the author's. Five substantive corrections produced by the meta-review process changed the project's claims; we report each in Section 6.

We do *not* claim that this methodology is the right one in every domain, that this run is representative of long-running agent behavior in general, or that the Voynich findings constitute decipherment, identify a source language, or rule out untested mechanism classes (transposition ciphers, running-key ciphers, position-sensitive ciphers of non-Latin source languages, structured glossolalia). We narrow the space of viable generative mechanisms for the manuscript; we propose specific falsifiable claims about agent methodology; and we present one full project run as evidence under both headings.

## 2 The Cryptanalyst Architecture

This section describes the methodology that produced the case study. We present the architecture as it actually ran rather than as an idealization; the failure modes it caught, and those that slipped through, are reported in Section 6.

### 2.1 Identity and the Loop

The primary agent is given an identity document defining its role (a Cryptanalyst investigating what the manuscript actually is), its goal (narrow the hypothesis space about generative mechanisms), and a 10-step iteration cycle: (1) read state, (2) select a hypothesis from the backlog, (3) check prior work in the knowledge base, (4) design the experiment, (5) write and run the script, (6) interpret the result, (7) record the finding via the write-finding skill, (8) update state files, (9) log the iteration, (10) repeat or pause if human judgment is required. Behavioral constraints in the same document include: never claim decipherment; treat positive results as having a high prior of being methodological artifacts; always include control corpora; distinguish exploratory analysis from hypothesis tests; keep experiments small (under 10 minutes runtime, single clear result, understandable in two paragraphs).

The constraints are not exhortations. They are the raw material the audit-experiment skill (§2.2) uses to generate explicit checklists run after every experiment, before any state-file update.

### 2.2 The Five Skills

The architecture provides five domain-tailored procedural skills, each instantiated as a markdown specification the agent reads before invoking the corresponding action.

**run_experiment.** Specifies the experiment lifecycle: directory structure (`experiments/exp_NNN_name/{script.py, result.md, data.json}`), pre-registration template, the mandatory "Expected Outcomes" section that must include an explicit statement of the form *"this experiment is an UNFAIR test of the hypothesis if [condition]"*, and a result template with required sections (Hypothesis, Method, Expected Outcomes, Raw Output, Result, Interpretation, Limitations, Self-Audit, Follow-up Questions). The unfair-test disclosure must be filled in before any code runs; filling it in afterward is explicitly prohibited.

**audit_experiment.** Specifies five checks to run on every draft result before any state-file update:

Table 1: Persistent state files in the Cryptanalyst architecture, with their roles and case-study sizes by project closure (exp_588).

| File | Role | Size at closure |
|---|---|---|
| `working_position.md` | Live theory-discrimination synthesis | ∼316 lines |
| `findings.md` | Established facts (F001–F642), grep by ID | 4,836 lines |
| `failed_approaches.md` | Ruled-out hypotheses (X001–) | — |
| `open_questions.md` | Items needing human judgment (Q001–) | — |
| `hypothesis_backlog.md` | Active untested hypotheses | — |
| `committed_claims.md` | Public claims (C001–C009); only the human edits | 9 entries |
| `proposed_commitments/` | Agent proposes; human approves | — |
| `working_model.md` | Topical reference characterization | — |
| `iteration_log.md` | One-line per experiment, append-only | 553 entries |
| `current_state.md` | Periodic prose snapshots of project status | — |
| `archive/` | Meta-reviews, resolved hypotheses, paper materials | — |

1. *Target-awareness:* any parameter, table, or mapping calibrated using knowledge of the target metric must be disclosed and the result stated as conditional on the calibration.
2. *Feature-destroying implementation:* any implementation detail that destroys or degrades the feature the hypothesis depends on makes the experiment unfair to the hypothesis; the conclusion must read "this hypothesis fails under this implementation," not "this hypothesis fails."
3. *Extrapolation beyond the tested grid:* claims of "structural," "no parameter setting," or "any mechanism of class X" are softened to the actual tested range unless backed by an explicit logical argument.
4. *Negative-result overreach:* the valid conclusion from a negative result is "this mechanism did not produce X under these conditions," not "X requires Y" or "no mechanism of class Z can produce X."
5. *Control-comparison interpretation:* a control behaving unexpectedly, especially one that exceeds the real data on a metric, is investigated before the comparison is interpreted.

Each check originated in a specific past failure mode of this project (§6). The skill explicitly disclaims that it does not catch novel failure modes; the meta-review agent (§2.5) remains the primary independent check.

**write_finding.** Specifies how completed experiments enter persistent memory. Findings go to `findings.md` as positive constraints, ruled-out hypotheses go to `failed_approaches.md`, and questions needing human judgment go to `open_questions.md`. Every entry must cite the source experiment ID (the *traceability rule*: no entry of the form "I noticed that…" or "it seems like…" without an experiment citation is permitted) and must include *magnitude context*: an explicit comparison to any published value for the same metric and to the closest prior finding in the project. The intent of magnitude context is operational: a finding $10\times$ larger than any published value on the same metric is either a breakthrough, an artifact, or a unit error, and the agent should notice which before moving on. The skill also defines the *contradiction rule*: before any state update, the agent checks the commitment ledger; if a hard contradiction exists with a committed claim, the loop stops and the issue is written to `open_questions.md` for human resolution.

**generate_hypotheses.** Specifies how the hypothesis backlog evolves: deriving 1–3 follow-ups after every experiment by asking what the result did not explain, what would falsify the interpretation, and what the control corpora revealed; recognizing when a finding opens a genuinely new direction; injecting novelty when the backlog is depleted (decompose the unit of analysis, swap the unit, import a method from a neighboring field); and phrasing hypotheses to be specific (naming a measurable quantity), falsifiable (with a possible refuting result), discriminating (favoring one theory over others), and bounded (testable in one experiment session).

**query_kb.** Specifies the procedure for searching the prior-literature knowledge base before designing an experiment, to avoid retesting hypotheses already settled in the field. Used at the start of every hypothesis evaluation cycle.

### 2.3 Persistent State

The architecture treats state-file curation as a first-class engineering concern rather than emergent behavior. Eleven distinct state files served different roles in the case study run.

The split between `findings.md` (the agent's internal record) and `committed_claims.md` (public-facing claims, only human-edited) is load-bearing. It separates the high-volume internal state, where the agent moves quickly and is allowed to be wrong, from the slow-moving public commitment surface that the agent is expected to defend across iterations.

The `proposed_commitments/` workflow makes the human's approval gate explicit: the agent drafts a proposal file, the human approves and applies it to the ledger, the proposal is moved to an applied subdirectory for traceability.

### 2.4 Housekeeping Cadence

A separate housekeeping document specifies cadenced cleanup actions that prevent state-file drift over long horizons:

- Every 10 experiments, resolved hypotheses move from the active backlog to an archive file.
- Every 30 experiments, a dated paragraph is appended to `current_state.md`.
- When the live synthesis document exceeds 500 lines, it is archived and rewritten; degradation from synthesis into append-only log is the failure mode this prevents.
- If 5 or more consecutive experiments target the same narrow topic, the agent switches category (word-internal structure, cross-unit patterns, anomalies, or simulation implications).

These cadences exist because, in their absence, autonomous loops accumulate state-file drift. We observed this drift directly during the case study (§6.4); the rules above were added or strengthened in response.

### 2.5 Multi-Agent Meta-Review

The primary agent (Sonnet-class) runs the loop continuously. A higher-capability meta-review agent (Opus-class) is invoked at phase boundaries: after the characterization phase, after the discrimination phase, after the architecture-construction phase, and at project closure. The meta-review reads a defined slice of state files and recent experiment results and produces a comprehensive review document flagging overstatements, target-awareness issues, scope claims that exceed evidence, and unflagged assumptions. Four such reviews were produced during the case study; each generated specific actionable corrections that the primary agent then implemented in subsequent experiments.

The audit-experiment skill (§2.2) is explicitly framed as load-reduction for the meta-review, not as a substitute. Its five checks catch known failure-mode categories before they reach the reviewer; the reviewer catches the rest, including novel categories the audit skill cannot anticipate.

### 2.6 Multi-Phase Project Design

The Cryptanalyst phase reported here was preceded by a Librarian phase that built a knowledge base of 100+ summarized prior-literature sources and produced a registry indexing them by topic, citation graph, and ingestion status. The Cryptanalyst's query-KB skill consults this knowledge base at the start of every hypothesis evaluation, to avoid retesting questions already settled in the field. The phase design is an architectural decision, not an emergent property; it makes knowledge curation and hypothesis testing distinguishable activities with separate evaluation criteria.

## 3 Hypotheses About Long-Running Research Agents

This section advances four hypotheses about when the architecture in Section 2 produces trustworthy work. We frame each as falsifiable. The case study supports each, but support is not proof; we mark the supporting evidence and the conditions that would refute the hypothesis. The hypotheses are offered to be tested in subsequent multi-domain work; we do not claim they are exhaustive. Three further architectural choices that may have contributed to the case study's outcomes but for which we have only single-run introspection are discussed separately in §9.1.

**H1 (External meta-review catches errors the primary agent's self-checking misses).** Even when the primary agent has explicit failure-mode checklists embedded in its skills, errors of those exact categories slip through; an independent higher-capability reviewer is needed to catch them.

*Evidence.* All three documented failure modes that motivated the audit-experiment skill in this project (Catches 1–3 in §6.1) were caught by the meta-review agent, not by the primary agent's self-checking. The audit skill was created *because* those failures occurred in self-audited experiments. Two additional corrections during the run (a misattribution in committed-claim text, a line-position convention check) were also caught externally. We are aware of no major correction in this project that the primary agent caught in self-audit before the meta-review surfaced it.

*Refutation condition.* A run in which the primary agent catches and corrects a comparable substantive error (target-awareness, feature-destroying implementation, negative-result overreach, or unit/attribution error) without external prompting, on a finding it had already submitted for state update.

**H2 (Commitment ledgers separate fast-moving findings from slow-moving claims).** A two-tier system in which the agent freely writes to an internal findings file but only the human edits a separate commitment ledger prevents claim drift over long horizons.

*Evidence.* Over 579 experiments, the internal findings file accumulated 642 entries; the commitment ledger held 9 entries throughout. A documented misattribution in the ledger (C007's "2.05× attributed to A `ok`" actually belonged to B `qo`) was caught during a ledger audit and amended through the propose-then-commit workflow rather than by direct edit. Without the ledger separation, drift between live characterization and public-claim text would not have been detectable as a discrete event.

*Refutation condition.* A run in which an agent without a commitment ledger nonetheless maintains identical alignment between internal findings and public claims over hundreds of iterations.

**H3 (Pre-registration with explicit "unfair-test" disclosure reduces overreach errors).** Forcing the agent to write down before running an experiment what implementation details would make the experiment unfair to the hypothesis catches a class of negative-result overreach that retroactive review does not.

*Evidence.* The exp_180 failure (Catch 3) was an experiment whose modification operation actively destroyed the feature the hypothesis depended on; the negative result was attributed to copy-buffer architectures in general; exp_181 reversed the conclusion by preserving onsets. The audit-experiment skill's Check 2 was created to enforce that this disclosure happen before the experiment runs, not after. After Check 2 was institutionalized, no analogous failure occurred in the remaining ∼400 experiments.

*Refutation condition.* A controlled comparison showing equivalent overreach rates between agent runs with and without the unfair-test pre-registration requirement.

**H4 (Magnitude context catches anomalous findings the agent would otherwise report normally).** A rule requiring every numeric finding to be compared to published values for the same metric and to the closest prior finding in the project catches unit errors and extraordinary claims early.

*Evidence.* The write-finding skill institutes this rule explicitly. The C007 misattribution (a 2.05× value attributed to the wrong dialect/class) was caught partly because the commitments audit treated the value as a claim to verify against its supporting experiment, exposing the inconsistency. The rule is also load-bearing for noticing the convergent-failure structure of the discrimination result (§5): both Naibbe and Timm fall an order of magnitude short of the $-am$ line-final target, which is interpretively important only if the agent is required to think about the magnitude.

*Refutation condition.* A controlled comparison showing equivalent unit-error and misattribution rates with and without the magnitude-context rule.

These four hypotheses are the original methodological contribution of this paper. The case study evidence is one project run; it supports the hypotheses but does not establish them. We present the case study in the next four sections both as substantive Voynich research and as the data on which the hypothesis evidence above rests.

## 4 Case Study (I): Voynichese Structural Grammar

This section reports the structural-grammar findings the agent produced over the characterization phase of the case study. Existing characterizations (Currier, 1976; D'Imperio, 1978; Stolfi, 2000; Zattera, 2022; Bowern and Lindemann, 2021) establish that Voynichese has a Currier A/B dialect split—two statistically distinct populations of folios first identified by Currier 1976—a productive suffix paradigm, a low-high-low entropy profile within words, and slot-grammar structure. Building on the agent's ∼579 corpus-statistical tests against the EVA ZL3b transliteration (Zandbergen, 2023)—EVA (Extended Voynich Alphabet) being the standard scholarly system for rendering the manuscript's script in Roman letters—we extend these in five directions.

A note on terminology used throughout this section. Voynichese words are grouped by *onset class*—the leading letter sequence in EVA notation (for example, words beginning *d-*, *sh-*, or *qok-*)—into roughly 20 functionally distinct categories that act as word-class proxies. *Positional enrichment* is a ratio: 1.0 means a class appears at a given line position exactly as often as chance predicts; 5.23× means more than five times as often. A result is *confirmed* (CONF) if it clears a conservative statistical threshold correcting for the large number of comparisons made (Bonferroni correction at $m$=175 tests, minimum 10 observations). A *null model* is a randomly shuffled version of the data used to establish what patterns could arise by chance alone. All five findings below were *pre-registered*—their methods and acceptance thresholds committed to a timestamped record before experiments ran—to guard against post-hoc adjustment.

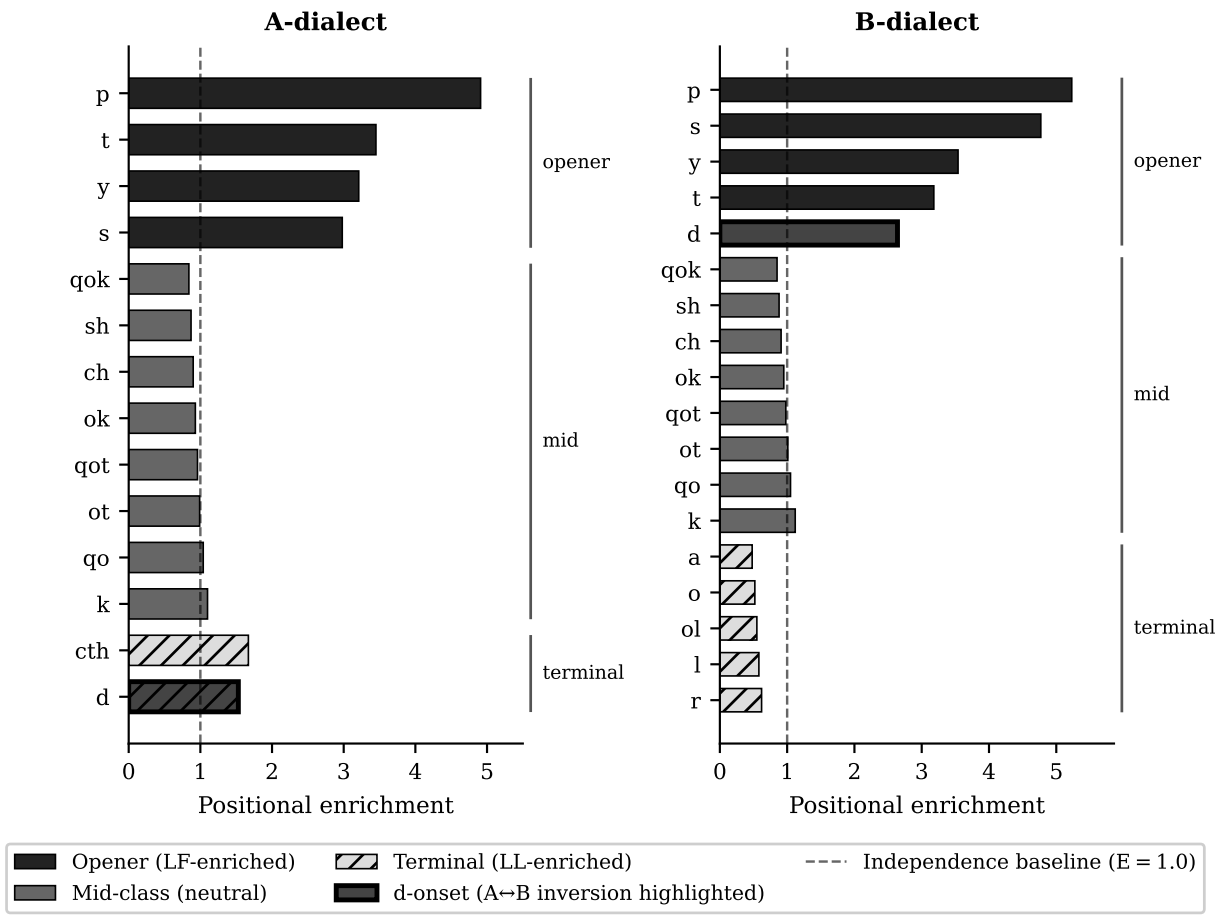


Figure 1: Three-layer positional architecture. Horizontal bars show positional enrichment for each onset class in A-dialect (left) and B-dialect (right). Openers (dark) are enriched at line-first position; terminals (hatched) are enriched at line-final position; mid-classes are neutral at both boundaries. The **d**-onset (highlighted) is the dominant B-dialect opener and the dominant A-dialect terminal—the sharpest cross-dialect inversion in the data.

**Three-layer positional architecture (Figure 1).** Onset classes partition by line-position enrichment into openers (confirmed as line-first enriched), mid-classes (neutral at boundaries), and terminals (confirmed as line-final enriched). The B-dialect openers are $\{p\,(5.23\times), s\,(4.77\times), y\,(3.54\times), t\,(3.18\times), d\,(2.64\times)\}$—meaning, for instance, that $p$-onset words appear at line-first position more than five times as often as chance would predict; mid-classes $\{ch, qok, ok, ot, qot, sh, k\}$; terminals $\{r\,(2.80\times), l\,(1.85\times), ol\,(1.65\times), o\,(1.58\times), a\,(1.52\times)\}$. The A-dialect partition is structurally analogous but with critical inversion: A terminals are $\{d\,(1.53\times), cth\,(1.67\times)\}$, sharing zero classes with B terminals. This terminal-layer inversion is the cleanest dialect discriminant in the manuscript: $d$-onset is the dominant A line-closer ($18.4\%$ of A lines) and the dominant B line-opener ($19.2\%$ of B lines). A complete glossary of shorthand used throughout ($ss_y$, $ss_d$, $r_{\text{lf}}$, $r_{\text{mid}}$, ZeroMid, $\beta_{\text{scale}}$, $am_{\text{ll}}$, and related terms) appears in supplementary material.

**Three-register line grammar.** Lines partition by folio position into openers (first line of each folio, $67$–$80\%$ gallows-onset LF, mean 7–8 words, $85\%$ hapax line-final vocabulary), continuations (baseline distributions), and closers (last line, anti-gallows LF $0.44$–$0.60\times$, mean 4–5 words). Gallows glyphs are the tall ascender characters in EVA ($p$, $t$, $k$, $f$); hapax words are those appearing only once in the corpus. The pattern is section-invariant. The A-dialect $p$-onset shows complete folio-position exclusion: $37.3\%$ of folio-opener line-final tokens, $0$ of $83$ folio-closer line-final tokens. The gallows-opener formula is a one-word constraint operating on word-index $0$ only, with no internal template structure—the line-final vocabulary is dominated by words appearing only once.

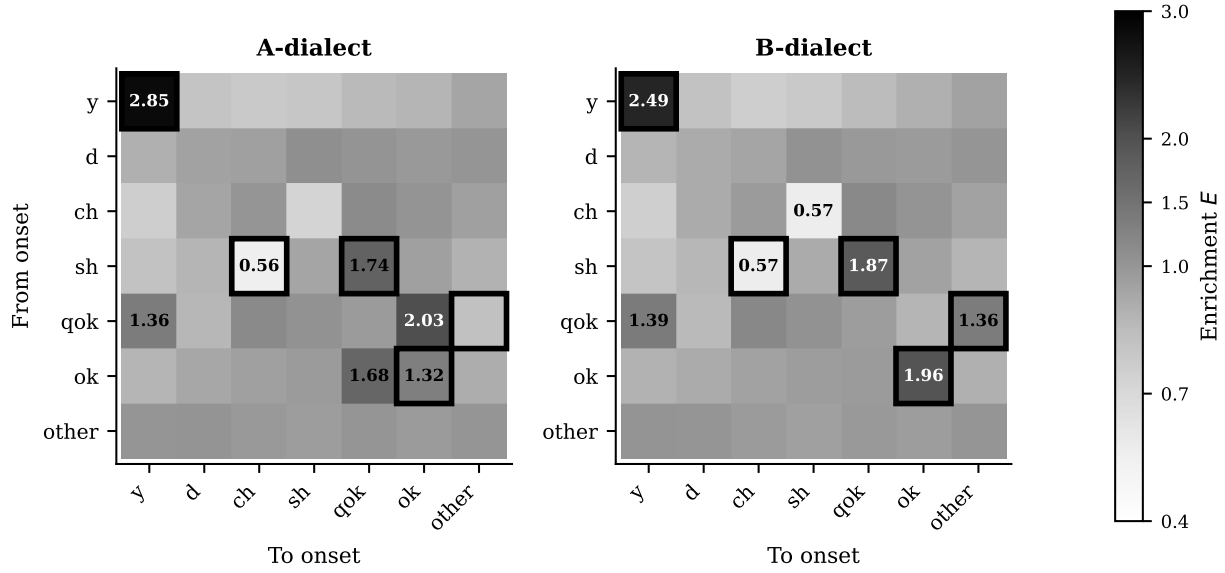


Figure 2: Onset-class bigram succession matrix for A-dialect (left) and B-dialect (right). Shading encodes enrichment $E$ (dark = elevated, white = depleted, mid-gray $\approx 1$); values are annotated where $E \geq 1.3$ or $E \leq 0.7$. Bold borders mark the five cross-dialect invariants: $y \to y$, $sh \to qok$, $ok \to ok$, $qok \to$other (B), and $sh \to ch$ depletion.

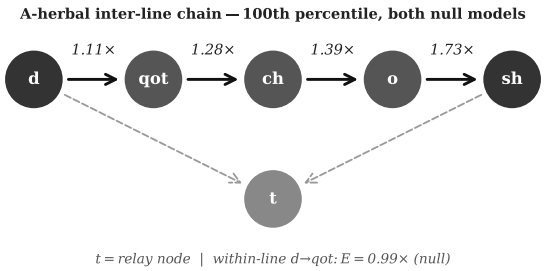


Figure 3: A-herbal paragraph-level chain ($d \to qot \to ch \to o \to sh$, with $t$ as relay). Edge labels show inter-line enrichment values. Note that within-line $d \to qot$ is null ($E = 0.99\times$); the chain operates strictly at the cross-line level. The chain lies at the 100th percentile of both section-shuffled and onset-shuffled null distributions ($0/200$ controls in each case).

**Bigram grammar with cross-dialect invariants (Figure 2).** A bigram is a pair of successive words; here we track which onset class tends to follow which other. The complete $7\times7$ onset-class succession matrix, computed relative to a chance baseline, yields five cross-dialect invariant elevated pairs—pairs that follow each other more often than expected in both dialects ($y \to y$ at $2.49\times/2.85\times$, $sh \to qok$ at $1.97\times/1.74\times$, $ok \to ok$ at $1.32\times/1.96\times$, $qok \to s$ at $1.36\times/1.39\times$)—and one cross-dialect avoidance, meaning the two classes follow each other far less often than chance ($sh \to ch$ at $0.555\times/0.565\times$). B-dialect adds bidirectional $ch \leftrightarrow sh$ avoidance ($ch \to sh = 0.567\times$). A-dialect adds $qok \leftrightarrow ok$ bidirectional coupling ($2.03\times/1.68\times$) absent in B. The $sh \to qok$ coupling is driven by the $she$- body subclass (*shedy*, *shey*, *sheedy*) operating as a triple-hub node receiving $d \to sh$ and $s \to sh$ and emitting $sh \to qok$.

**Paragraph-level chain grammar (Figure 3).** A herbal sections exhibit a directed acyclic cross-line succession network: $d \to qot \to ch \to o \to sh$, with $t$ as a relay class receiving from $d$ and $sh$. The chain operates strictly at the inter-line level: within-line $d \to qot$ succession is null ($A : 0.99\times$, $B : 0.66\times$ depleted), while inter-line $d \to qot$ is CONF. This is the clearest scope-separation signal in the data—two distinct mechanisms operate at different structural levels. The chain is A-herbal-specific; B-dialect cross-line patterns are heterogeneous and section-specific (B astro: $qok$

Table 2: Transliteration replication of Section 4 findings. Retained/Failed per pre-registered thresholds (committed before experiments). Finding 4 under v101 falls below the 95th-percentile null-model threshold due to reduced observation counts on one chain edge.

| Finding | EVA-basic | v101 | ZL3a | Overall |
|---|---|---|---|---|
| F1: Three-layer architecture | 18/18 R | 16/18 R | 18/18 R | **Retained** |
| F2: Terminal inversion | R | R | R | **Retained** |
| F3: Bigram invariants | 4/5 R | 4/5 R | 4/5 R | **Retained** |
| F4: Paragraph chain | R | **F** | R | **Retained (2/3)** |
| F5: Suffix transformation | 20/21 R | 19/22 R | 20/21 R | **Retained** |

self-chains; B pharma: $ol{\rightarrow}qok$ at $2.76\times$). Under null-model testing—that is, comparing the chain against 200 versions of the data with sections randomly shuffled and 200 versions with onset labels randomly shuffled—0 of 200 controls in either case reproduce the chain's joint per-edge strength. The chain lies at the 100th percentile of both null distributions. We report the chain as a theoretically-motivated directed finding specified a priori by the bigram characterization in the previous paragraph, not as a globally-strongest chain identified by exhaustive search; this distinction matters because exhaustive-search statistics over all directed length-5 paths are dominated by sparse-class pathology, while the specified chain survives the proper null.

**Systematic dialect transformation.** Across every onset class, A-dialect uses $-ol$/$-or$/$-ey$ suffixes and B-dialect uses $-edy$/$-eedy$. This is the most systematic vocabulary transformation between dialects, affecting $20+$ characterized onset classes uniformly. Combined with the terminal-layer inversion and the structural identity $\text{adj\_rate}_A = \text{adj\_rate}_B = 0.00933$ (matching to four decimal places), this points to systematic encoding difference rather than scribal drift.

**Transliteration replication.** All five characterization findings were replicated under three alternative transliterations (EVA-basic, v101 Claston, and ZL3a) against pre-registered thresholds committed to the project repository before the experiments ran. Table 2 reports the per-finding, per-transliteration results. Four of five findings replicated in all three systems; the A-herbal paragraph chain (Finding 4) replicated under EVA-basic (96th/97.5th percentile) and ZL3a (97.5th/96.5th percentile) but fell below the pre-registered 95th-percentile threshold under v101 (77th/86.5th percentile), attributable to reduced observation counts on the $o{\rightarrow}sh$ chain edge under v101's tokenization (obs $= 1$ vs obs $= 4$ under ZL3b). Per the pre-registered rule requiring retention in $\geq 2$ of 3 alternative transliterations, all five findings are classified as retained.

A separate finding bears on interpretive claims about specific Voynich words. The token *daiin* accounts for $27.4\%$ of all $d$-onset tokens. Its dominance is fully explained by the convergence of $d$-onset frequency, the $d$-onset $\rightarrow -aiin$ body constraint ($34.6\%$ A, $22.5\%$ B), and its position-neutral status (ZeroMid: enriched at neither line boundary)—it requires no special semantic explanation. Interpretations treating *daiin* as a key content word should account for this mechanistic baseline.

## 5 Case Study (II): Mechanism Discrimination and Constructive Existence

This section reports the discrimination phase. We test specific generative mechanisms representing the three main hypotheses about Voynichese—natural-language cipher, mechanical procedure, structured hybrid—and find that no position-blind or substitution-only mechanism reproduces the joint structural profile established in Section 4. We then construct one hybrid that does, as evidence the surviving mechanism class is non-empty.

### 5.1 Mechanisms tested

We evaluate generative mechanisms against the Currier B corpus (2,447 lines, 20,490 tokens) under matched line-length distributions. Reported confidence intervals are empirical 2.5–97.5 percentiles.

**Naibbe** (Greshko, 2025): a homophonic substitution cipher in which each plaintext character can map to multiple possible ciphertext characters, recently proposed as a specific generative model for Voynichese. Tested at published parameters, applied to Italian (Dante) and Latin (Vulgate) plaintexts under two line-breaking protocols (400 runs).

**Timm self-citation** (Timm and Schinner, 2020): an algorithm that generates new text by copying and lightly modifying words from nearby positions in the growing output, proposed as a mechanism that can produce Voynich-like word distributions without encoding any natural language. Tested across a 75-point parameter sweep (COPY_PROB $\in \{0.3, 0.5, 0.7, 0.85, 0.95\}$, MODIFY_PROB $\in \{0.1, 0.25, 0.4, 0.6, 0.8\}$, WINDOW $\in \{10, 30, 50\}$; 1,500 runs).

**Position-sensitive substitution cipher**: a cipher that applies different character substitution tables depending on whether a word falls at the start, middle, or end of a line, designed to test whether position-dependent rules can reproduce Voynichese's positional structure. Applied to Cicero *De Officiis*. We disclose that the line-final table is target-aware: all Latin clause-final morphemes ($-am$, $-um$, $-us$, $-em$, $-is$, $-ae$, $-orum$, $-bus$, $-unt$, $-it$, $-ant$, $-at$) map to EVA $-am$ specifically, because $-am$ is the metric target. This makes the resulting null finding stronger, not weaker.

**Hybrid architectures.** Four configurations: position-blind copy buffer over the cipher (exp_179[1]); position-aware buffer restricted to mid-line with onset-replacing modification (exp_180); position-aware buffer with onset-preserving suffix-replacement modification (exp_181); position-aware buffer with Timm-style character-level modification preserving onset (exp_182). Total: $4{,}400+$ hybrid simulations.

**A position-aware copy-buffer hybrid with full calibration.** A 21-mechanism variant in which line-first and line-final words are generated by a position-sensitive onset/suffix mechanism, mid-line words are drawn from a rolling copy buffer with onset-preserving character-level modification, and additional mechanisms perform vocabulary clamping; Markov onset coupling (in which each onset class's probability depends on the preceding class, reproducing the bigram structure from Section 4); $n$-conditional Menzerath scaling (an empirical linguistic pattern in which longer words tend to occur alongside shorter ones, applied here to match Voynich word-length distributions); and short-line line-final word-length filtering. Six parameters are explicitly target-calibrated. Full enumeration of mechanisms and parameter values appears in supplementary material. This variant is treated as a constructive existence proof (§5.3) rather than as a further mechanism under discrimination test.

### 5.2 The dual constraint

Two metrics drive the discrimination. *Self-succession* measures how often a given onset class follows itself on adjacent words relative to chance: $1.0$ means as often as chance predicts; $1.47\times$ means $47\%$ more self-following than expected. The *adjacent identical word rate* (adj_rate) is the fraction of consecutive word pairs that are exact repeats. $-am$ *line-final enrichment* is the positional enrichment of words ending in the $-am$ suffix at the last position of a line—Voynich B's strongest positional signal.

**Position-blind mechanisms fail on positional grammar.** Naibbe produces $-am$ line-final enrichment of $1.008\times$ (vs. Voynich $6.811\times$); the Timm sweep maximum across $75$ combinations is $0.572\times$. The convergent failure of two architecturally distinct mechanisms at the same metric establishes that we are ruling out the *class* of position-blind mechanisms, not specific implementations.

**Substitution alone produces no sequential memory.** The target-calibrated position-sensitive cipher reaches $5.756\times$ on $-am$ line-final enrichment ($84\%$ of Voynich) but produces self-succession of exactly $0.999\times$—the independence expectation. This is not a calibration issue; substitution ciphers operate token-wise.

**Timm produces sequential memory at wrong magnitudes.** Across the parameter sweep, no setting jointly produces Voynich-level self-succession ($\sim 1.47\times$) and Voynich-level identical-adjacent rate ($0.009$). At published parameters, Timm produces self-succession $1.5$–$2.4\times$ but adj_rate $0.037$ (fourfold excess). Reducing COPY_PROB lowers both metrics in lockstep.

**Hybrid progression.** The position-blind hybrid (exp_179) fails because copying at line-final positions dilutes the positional signal linearly: each $+0.1$ in COPY_PROB costs $-0.47$ in $-am$ line-final enrichment. Position-aware gating (exp_180) decouples the positional signal ($-am$ line-final enrichment flat at $5.7$–$5.8\times$ across all COPY_PROB) but onset-replacing modification destroys the self-succession feature, producing $0/18$ jointly viable combinations. Suffix-replacement modification (exp_181) decouples all three metrics and reaches $4/36$ jointly viable, with best-combination adj_rate $0.0144$ ($1.6\times$ above Voynich). Character-level modification preserving onset (exp_182) reaches $3/36$ viable at the strict adj_rate $\leq 0.009$ threshold: the first mechanism to cross the Voynich exact-repetition rate while maintaining the positional and sequential signals.

Table 3 and Figure 4 summarize. The dual constraint defines minimum requirements for any viable generative hypothesis among the mechanism classes tested; the position-aware copy-buffer hybrid with morphologically plausible modification is shown by exp_182 to contain at least one satisfying configuration.

[1]Experiment identifiers (exp_*NNN*) reference per-experiment results in supplementary material.

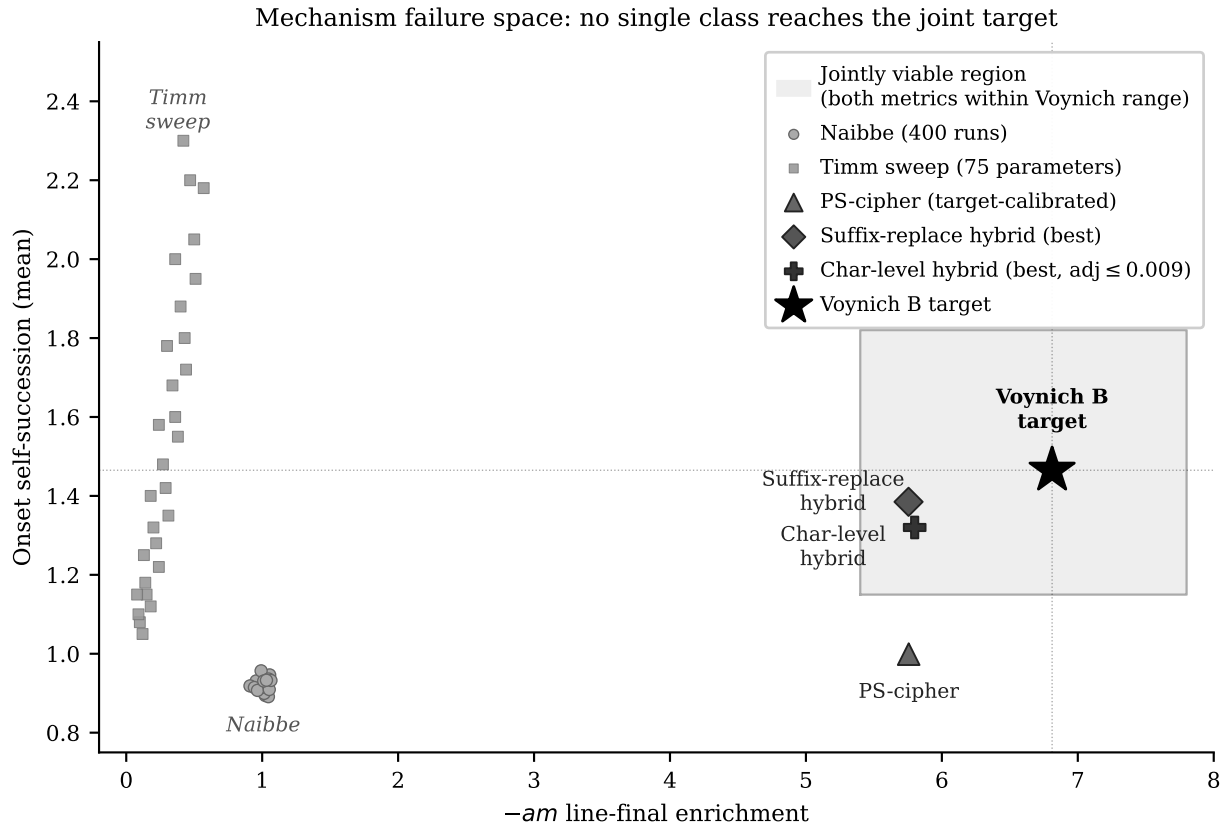


Figure 4: Mechanism failure space. Each tested mechanism class is plotted by $-am$ line-final enrichment (x-axis) and onset self-succession mean (y-axis). The Voynich B target (star) requires satisfying both axes simultaneously. The gray shaded region marks the jointly-viable zone. No position-blind cipher (Naibbe, Timm) approaches the x-axis target; substitution alone (PS-cipher) fails the y-axis; hybrid progression moves monotonically toward the target as architectural constraints are added.

Table 3: Mechanism failure matrix. **Bold** indicates the value falls outside the Voynich-consistent range. PS-cipher uses the target-calibrated strong setting.

| Metric | Voynich B | Naibbe | Timm (sweep) | PS-cipher |
|---|---|---|---|---|
| $-am$ line-final enr. | 6.811 | **1.008** | **0.572** | **5.756** |
| $y$-LF | 3.506 | **1.011** | **0.526** | 2.395 |
| Self-succession (mean) | 1.465 | **0.92** | 1.5–2.4 | **0.999** |
| Adj. rate | 0.009 | $\approx 0.009$ | **0.037** | $\approx 0.009$ |

### 5.3 Constructive existence

The discrimination result narrows viable mechanisms, among the mechanism classes tested, to a position-aware copy-buffer hybrid class with onset-preserving modification; other mechanism classes not evaluated here may also satisfy the joint constraints. To establish that this class is non-empty at the full joint profile, we construct one such hybrid and verify it reaches all 19 metrics of a B-dialect battery within $\pm 20\%$ tolerance under target calibration ($N$=30). Structurally, the hybrid generates line-first and line-final words by position-sensitive onset/suffix mechanisms and mid-line words by a rolling copy buffer with onset-preserving character-level modification; supplementary mechanisms perform vocabulary clamping; Markov onset coupling ($p_{qok \to ch} = p_{ch \to qok} = 0.12$); and $n$-conditional Menzerath scaling. Six parameters are target-calibrated; the remainder are structural. The construction demonstrates that the dual constraint is satisfiable—that some member of the surviving class attains the joint profile—but is not intended as evidence that this mechanism class is the generative model of Voynichese.

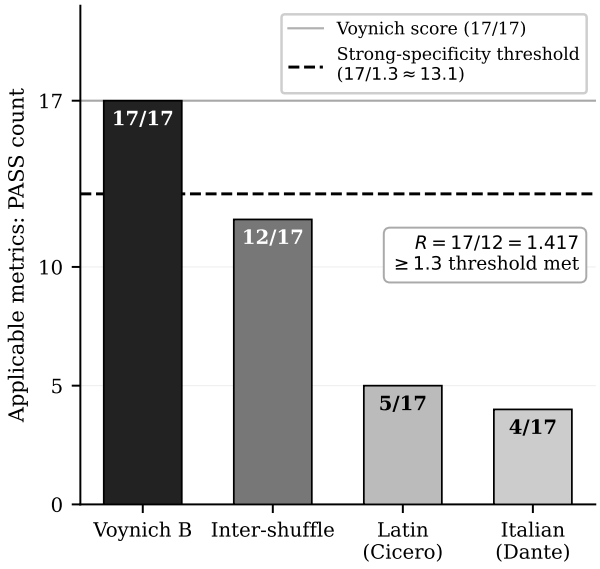


Figure 5: Architecture specificity: applicable battery PASS counts across four test corpora. The dashed line marks the pre-registered threshold below which non-Voynich corpora must fall for strong specificity ($17/1.3 \approx 13.1$). All non-Voynich corpora fall below this threshold; the specificity ratio $R = 17/12 = 1.417$ meets the pre-registered $R \geq 1.3$ criterion.

Degrees of freedom should be disclosed explicitly. The architecture has 21 component mechanisms with 6 target-calibrated parameters against a 19-metric battery whose effective independent dimension is lower than 19. Under PCA on the cross-simulation metric covariance, 16 components capture $95\%$ of variance; under hierarchical correlation clustering at $|r| \geq 0.7$, the battery resolves to 18 independent clusters. One pair is mechanically linked at $r = -0.94$ ($am_{\text{ll}}$ and $am_{\text{mid\_enr}}$, jointly determined by the line-final suffix mechanism). Even at the more favorable reading (18 effective constraints against 6 calibrated parameters), the architecture is not overdetermined, and we do not claim its $19/19$ battery success is itself evidence for the mechanism class. Under a pre-registered specificity test (exp_585; thresholds fixed before results were observed), the architecture was re-optimized against two structurally-matched non-Voynich corpora (Latin Cicero *De Officiis* and Italian Dante *Divina Commedia*, both segmented to Voynich-matched line-length distributions) and against inter-shuffled Voynich. The architecture achieves $17/17$ applicable metrics on real Voynich, $5/17$ on Latin, $4/17$ on Italian, and $12/17$ on inter-shuffled Voynich, yielding a pre-registered specificity ratio $R = 17/12 = 1.417$ (threshold $R \geq 1.3$ for strong specificity; met).

The construction yields one falsifiable mechanistic finding (C009). The architecture overproduces adjacent identical word pairs by 1.4–1.8$\times$ for *sh*/*ch*/*d* classes independent of vocabulary clamping. Diagnosis: Voynich adj_rate for these classes is consistent with random draws from finite vocabulary pools, and Voynichese has no special anti-repetition mechanism for them; the overproduction arises from copy-buffer/self-succession coupling. The resulting claim, that any copy-buffer with $\geq 8\%$ same-onset self-succession overproduces adj_rate for that class independent of vocabulary calibration, further constrains the surviving mechanism class beyond what the discrimination alone establishes. Separately, *s*-onset and *t*-onset show adj_rate $= 0$ in Voynich B ($0/452$, $0/466$ tokens), a structural zero from their positional constraint as line-openers.

A section-aggregation caveat applies. The B-dialect corpus pools four sections with systematically different positional grammar profiles: bio ($am_{\text{ll}}$=0.053), herbal, astro, pharma ($am_{\text{ll}}$=0.192). Document-order metric drift (first-half $am_{\text{ll}} = 5.825$, second-half 7.147) is 97–107$\%$ explained by section composition. The $19/19$ is therefore a full-corpus aggregate; section-level batteries are the appropriate next test and have not been performed.

## 6 What the Architecture Caught

This section reports the substantive corrections produced during the case study. They are the empirical evidence for the methodology hypotheses in Section 3: each correction names something the architecture caught (or, in one case, failed to catch promptly), and the scaffolding component that made the catch possible.

### 6.1 Three failure modes that motivated the audit-experiment skill

The audit-experiment skill (§2.2) did not exist at project start. It was created in response to three early failures, each caught by the meta-review agent and each becoming one of the skill's five checks.

**Catch 1 (Sampling artifact as signal; exp_175, mutual-information at lags 10–12).** The agent computed mutual information between words at increasing lag distances, comparing real Voynich to a within-line shuffle control. At lags 10–12, the within-line shuffle exceeded the real data, which the agent interpreted as evidence of real-data depletion at those lags, hence "long-range structure." The meta-review identified that the within-line shuffle's behavior at lags 10–12 was a sampling artifact from sparse long-line data, not a meaningful baseline. The original "long-range structure" finding was retracted. *Generalized rule:* when a control behaves unexpectedly, especially when it exceeds the real data, investigate the control before interpreting the comparison. This became Check 5 of the audit skill.

**Catch 2 (Target-awareness without disclosure; exp_177, position-sensitive cipher).** The agent built a position-sensitive substitution cipher with separate substitution tables for line-first, mid-line, and line-final positions. The line-final table mapped all Latin clause-final morphemes ($-am$, $-um$, $-us$, $-em$, $-is$, $-ae$, etc.) to EVA $-am$ specifically, because $-am$ was the Voynich B target metric the cipher was being evaluated against. The result, "cipher reaches $5.756\times$ on $-am$ line-final enrichment ($84\%$ of Voynich)," was reported without disclosing that the table was calibrated on knowledge of the target. The meta-review caught the omission; the result was restated as conditional on the calibration. *Generalized rule:* any parameter, table, or mapping set using knowledge of the target metric must be named, and the result must be stated as conditional. This became Check 1.

**Catch 3 (Feature-destroying implementation; exp_180, onset-replacing modification).** The agent built a position-aware hybrid in which a copy-buffer's modification operation replaced the onset class of copied tokens. The result was $0/18$ jointly viable parameter combinations, interpreted as "copy-buffer architectures cannot decouple onset-class succession from exact-word repetition." The meta-review identified that the modification operation *actively destroyed* the feature (onset self-succession) the hypothesis depended on. Replacing the onset of every modified copy meant that any onset-class succession could come only from unmodified copies, which were exact-word repeats. exp_181 reversed the conclusion by preserving onsets: $4/36$ jointly viable. *Generalized rule:* before attributing a negative result to the mechanism class, verify that the implementation does not actively destroy the feature under test. This became Check 2.

The audit skill was authored after these three. It runs after every experiment, before any state-file update. After its institution, no analogous failure was caught externally in the remaining $\sim$400 experiments. This is consistent both with the skill being effective and with the meta-review agent's attention being elsewhere; a controlled comparison would be needed to distinguish.

### 6.2 The C007 misattribution

A commitments audit at exp_573 found that the value $2.05\times$ in committed claim C007 ("onset self-succession range") was attributed to "$ok$, A-dialect." The supporting experiment (exp_164) showed B $ot{\to}ot = 2.05\times$; A $ok$ self-succession was actually $1.42\times$. The misattribution had survived in committed-claim text since the claim was authored. Two further errors in the same claim were identified at the same time: the stated range ($1.24$–$2.05\times$) was wrong (canonical range from re-derivation is $0.78$–$3.86\times$), and a sub-claim ("residuals $\geq 1.2\times$ for all") failed for B $qok$/$ch$/$sh$. The corrections were drafted as a proposed amendment, reviewed, and applied through the propose-then-commit workflow. *Mechanism that caught it:* the commitments audit is a periodic verification step that re-checks every committed-claim headline number against its supporting experiment's result file. The split between fast-moving findings and the slow-moving commitment ledger (H2) made the audit possible; it would have no analogous workflow in a flat single-state-file architecture.

### 6.3 The line-position convention check (exp_581)

During a specification audit, the agent re-derived four high-stakes grammar-spec values from the raw corpus to verify that the position convention (LF = word index $0$; LL = word index $n-1$) had been applied consistently. An initial run with one tokenizer produced a $y$-onset LF value of $3.121\times$, $11\%$ off the committed $3.506\times$; correcting to the experiment's actual tokenizer recovered $3.478\times$ ($0.79\%$ deviation, within tokenizer noise). The audit confirmed the convention had not been silently inverted in any spec entry. *Mechanism that caught it:* explicit cross-check experiments authored as part of paper-preparation. The traceability rule (every finding cites an experiment ID) made it possible to re-derive the value from the original script.

### 6.4 Loop drift in the housekeeping audit

A loop audit conducted at exp_287 found that iteration-log entries had stopped at exp_199 and were missing for the next 88 experiments before the gap was noticed. `current_state.md` had been unmaintained for 106 experiments. `working_position.md` had grown to 1,388 lines as an append-only log rather than being rewritten as synthesis. The hypothesis backlog held 937 lines of resolved-but-unmoved entries. None of these had impaired any committed claim or produced a contradicted finding (the experimental work itself was sound), but the state files had drifted to the point where a researcher picking the project up cold would have struggled to orient. The audit produced explicit cadence rules (every 10 experiments archive resolved hypotheses; every 30 append a paragraph to current_state; rewrite working_position when it exceeds 500 lines) and a one-time cleanup. A second housekeeping audit at exp_574 confirmed the cadences had held. *Mechanism that caught it:* a periodic audit performed by the agent itself, but only after the gap had grown large enough to be conspicuous. This is a partial failure of the methodology: the audit ran only after the drift became severe, not as a continuous check. We discuss the cadence-rule response in §9.1.

## 7 Methods

This section reports Voynich-specific methodological detail. The methodology of the autonomous loop is in Section 2.

We use the EVA ZL3b transliteration (Zandbergen, 2023), filtering to paragraph-type loci, with dialect assignment by per-line $-edy$ rate (canonical methodology; supersedes earlier folio-level assignment). Yielding B-dialect: 40 folios, 2,447 lines, 20,490 tokens. Positional enrichment uses one-sided binomial exact tests, Bonferroni-corrected at $m$=175. Self-succession is computed as $P(C \mid \text{prev}{=}C)/P(C)$ over in-line adjacency pairs, with stratified position control. Cross-validation is $80/20$ stratified by dialect plus leave-one-section-out. Null-model tests for the A-herbal paragraph chain use 200 section-shuffled and 200 onset-shuffled controls, comparing the joint per-edge enrichment of the specified chain against the null distribution. Code, data, per-experiment directories, and random seeds are available on request; canonical methodology, an executable validator, the $19{\times}19$ metric-correlation matrix referenced in Section 5.3, a glossary of shorthand, and the full enumeration of the 21 architecture mechanisms are included as supplementary material.

Replication thresholds for the transliteration study (Section 4) and specificity thresholds for the architecture test (Section 5.3) were pre-registered in a timestamped repository commit before the respective experiments ran. The pre-registration document, commit hash, and deviation log are in supplementary material.

## 8 Limitations

### 8.1 Limitations of the Voynich findings

**Untested mechanism classes.** The mechanism discrimination rules out tested classes only: position-blind ciphers (Naibbe, Timm) and substitution-only ciphers (the position-sensitive cipher). Transposition ciphers, running-key ciphers, book ciphers, position-sensitive ciphers of non-Latin source languages, and structured glossolalia remain untested.

**Transcription dependence.** All five Section 4 findings were replicated under three alternative transliterations; four of five replicated in all three systems. Findings at the suffix-class level remain sensitive to transliteration choice at the margin.

**Line-as-unit assumption.** Positional grammar metrics assume physical lines correspond to linguistic units. This is shared with the literature but unestablished.

**Section heterogeneity.** All B-dialect aggregate findings are subject to the section-composition caveat in Section 5.3. Scribal-hand assignments (Davis, 2020) are near-perfectly co-linear with section composition in B-dialect, preventing independent attribution of the gradient to scribal variation.

**Architecture degrees of freedom.** The constructive-existence architecture has 21 component mechanisms and 6 target-calibrated parameters against a 19-metric battery with effective independent dimension of 16–18. Its $19/19$ PASS is constructive existence, not evidence of the correct generative model. Specificity against other corpus types (medieval Hebrew, Arabic, constructed languages with similar morphological complexity) remains untested.

**No decipherment, no source language, no unique mechanism identification.** Multiple mechanism families could potentially satisfy the same constraints.

### 8.2 Limitations of the methodology evaluation

**One project, one domain.** All evidence in Section 3 comes from a single Voynich case study. The hypotheses are framed as falsifiable claims to be tested in subsequent multi-domain work; the evidence here supports them but does not establish them.

**No controlled comparison.** We have no within-project counterfactual run without the meta-review agent, without the commitment ledger, without the audit-experiment skill, or with a generic research-agent prompt. Each hypothesis names a refutation condition but the case study cannot itself supply the comparison.

**Selection bias on reported corrections.** The five corrections reported in Section 6 are the ones that were caught. We have no enumeration of errors that were not caught (the unknown-unknown class). The audit-experiment skill explicitly disclaims novelty coverage; the meta-review may also have missed novel failure modes.

**Methodology under-evaluated.** Five caught errors in one project is encouraging but not a basis for strong claims about the autonomous-loop methodology. We intend to document the methodology separately in subsequent work with multi-domain runs designed for controlled comparison.

**Author-as-architect.** The author designed the identity document, the five skills, the commitment-ledger workflow, and the housekeeping cadence. The methodology evaluation is therefore not blind to its designer's intuitions about what should work; an independent re-instantiation by another researcher would be a stronger test.

## 9 Discussion: Generalizing the Methodology

The architecture in Section 2 was instantiated for a single domain. We conclude by considering where it would and would not transfer.

**Where the architecture should generalize.** Domains with the structural shape of the Voynich case—a clear empirical substrate, a substantial existing literature to anchor against (so rediscovery is detectable), the possibility of pre-registering falsifiable hypotheses, and a high prior on positive results being artifacts—are good candidates. Cryptanalytic challenges with held-out test sets, mathematical conjecture exploration where small-scale computation can refute conjectures, physics anomaly investigation where simulation can constrain mechanisms, and data-driven biological hypothesis generation against existing genomic or proteomic corpora all share enough of this shape that the same scaffolding (loop, skills, state files, commitment ledger, meta-review) could plausibly transfer.

**Where it would not.** Domains where ground truth is human judgment rather than empirical measurement (literary interpretation, value-laden policy questions, qualitative social-science synthesis) lack the falsifiability the audit-experiment skill's checks rely on; the skill cannot meaningfully evaluate "feature-destroying implementation" if the feature is not operationally definable. Domains requiring physical experiments (wet-lab biology, materials synthesis) cannot run the inner loop autonomously without robotic infrastructure outside the scope of the agent. Domains where experiments are individually expensive (large-scale clinical trials, particle-physics runs) violate the "keep experiments small, $< 10$ minutes" constraint that the loop's high-throughput discipline depends on.

**What is portable versus Voynich-specific.** The five-skill structure, the commitment-ledger workflow, the propose-then-commit human gate, the mandatory pre-registration with unfair-test disclosure, the magnitude-context rule, the traceability rule, the periodic meta-review, and the housekeeping cadences are domain-portable. The specific content of the audit-experiment skill (target-awareness about VMS metrics, the specific feature-destruction failure mode of exp_180) is domain-specific and would need to be re-derived from each domain's observed failure modes. We expect that any new domain instantiation would acquire its own handful of caught failures in early operation, each becoming a check in that domain's audit skill. The architecture's value is that this acquisition is structured: failures are documented, generalized into checks, and inserted into the skill so the same failure does not recur.

**What the case study does not tell us.** It does not tell us whether the methodology produces *better* findings than a researcher working alone over an equivalent wall-clock period would have. It does not tell us whether the per-experiment quality is higher or lower than that of a human experimenter. It does not tell us whether the meta-review's catches are representative of the catch rate that would be observed across many runs, or whether this run was unusually error-prone. These are the controlled-comparison questions that follow-on work needs to answer.

### 9.1 Architectural choices we suspect contributed but cannot demonstrate

Three further architectural choices were part of the case study and may have contributed to its outcomes, but we have only single-run case-study introspection to support them and so do not advance them as falsifiable hypotheses. We name them so that a reader designing a similar system can decide whether to inherit each.

*Domain-tailored skills versus a generic research-agent prompt.* The five skill files in this project were short (<10 KB each) and authored against domain-specific failures rather than generic research practice. We suspect this was a strength: the audit-experiment skill's checks each correspond to a specific past Voynich-domain failure, the write-finding skill's magnitude-context rule is operational rather than aspirational, the query-KB skill is specific to this project's knowledge base structure. But we have no side-by-side comparison against a generic research-agent prompt operating on the same hypothesis backlog, and cannot rule out that a generic prompt with a long enough context window would have caught equivalent errors.

*Cadenced housekeeping rules.* A loop audit at exp_287 found that iteration-log entries had stopped at exp_199 (an 88-experiment gap), `current_state.md` had been unmaintained for 106 experiments, the live synthesis document had degraded to a 1,388-line append-only log, and the hypothesis backlog had accumulated 937 lines of resolved-but-unmoved entries. After the audit, explicit cadence rules were added (every 10 experiments archive resolved hypotheses; every 30 append a paragraph to current_state; rewrite the synthesis document when it exceeds 500 lines). A second audit at exp_574 found the cadences had held. We suspect cadenced rules prevent state drift, but we have only a one-shot before/after comparison within the same project; whether the post-audit improvement was driven by the cadence rules or by renewed researcher attention is not separable.

*A separate prior knowledge-curation phase.* The Cryptanalyst phase was preceded by a Librarian phase that built a 100+-source knowledge base, with mandatory consultation at every hypothesis selection. We suspect this reduced wasted experiments on rediscovery, but we cannot quantify the counterfactual. The early characterization phase (exp_001–165) substantially rediscovered results already in the literature; without the knowledge base this fraction would plausibly have been higher, but "plausibly higher" is not evidence.

A reader designing a similar system should treat these three as live design choices rather than as architectural givens.

## 10 Conclusion

This paper has two contributions. The first is an architecture for autonomous long-running research agents (a multi-skill loop, persistent state files, a commitment ledger separating internal findings from public claims, mandatory pre-registration with explicit "unfair-test" disclosure, and a higher-capability meta-review agent invoked at phase boundaries) together with four falsifiable hypotheses about when the architecture produces trustworthy work. The second is a case study supporting those hypotheses: a 579-experiment autonomous investigation of the Voynich Manuscript that produced five replicated structural-grammar findings (a three-layer positional architecture with complete A/B-dialect terminal-layer inversion, a $7\times7$ bigram succession matrix with cross-dialect invariants, a directed paragraph-level succession chain at the 100th percentile of two independent null models, a systematic dialect-marking suffix transformation, and a mechanism-discrimination result establishing a dual structural constraint), all replicated under three alternative transliteration systems against pre-registered thresholds. The discrimination rules out position-blind and substitution-only mechanisms among the classes tested; a position-aware copy-buffer hybrid with onset-preserving character-level modification reaches the joint profile and yields one falsifiable mechanistic finding (C009: any copy-buffer with $\geq 8\%$ same-onset self-succession overproduces adjacent identical words for that class). Architecture specificity to Voynichese is confirmed by a pre-registered test against Latin and Italian corpora ($R = 1.417$, threshold $\geq 1.3$). Five substantive corrections caught during the run are reported in detail as evidence for the methodology hypotheses. Whether the architecture generalizes beyond this case study, across domains and across instantiations by other researchers, is the question subsequent work needs to answer.

## References

Claire Bowern and Luke Lindemann. The linguistics of the Voynich manuscript. *Annual Review of Linguistics*, 7:285–308, 2021.

Prescott H. Currier. Some important new statistical findings. Voynich manuscript conference proceedings, 1976.

Lisa Fagin Davis. How many glyphs and how many scribes? Digital paleography and the Voynich manuscript. *Manuscript Studies*, 5(1), 2020.

Mary E. D'Imperio. *The Voynich Manuscript: An Elegant Enigma*. National Security Agency, 1978.

Michael Greshko. The Naibbe cipher: a substitution cipher that encrypts Latin and Italian as Voynich Manuscript-like ciphertext. *Cryptologia*, 2025.

Jorge Stolfi. A grammar for Voynichese words. Technical report, Universidade Estadual de Campinas, 2000.

Greg Hodgins et al. Radiocarbon dating of the Voynich Manuscript. University of Arizona AMS Laboratory, 2011.

Torsten Timm and Andreas Schinner. A possible generating algorithm of the Voynich manuscript. *Cryptologia*, 44(1):1–19, 2020.

René Zandbergen. The EVA transliteration alphabet. `http://www.voynich.nu/`, 2023.

Massimiliano Zattera. A new transliteration alphabet brings new evidence of word structure and multiple languages in the Voynich manuscript. International Conference on Historical Cryptology, 2022.